Journal of Machine Learning and Information Security
https://www.sciltp.com/journals/jmlis

Article

MG-CMIF: A Multi-Granularity Enhanced Cross-Modal Information Fusion Framework for Molecular Property Prediction

Maoyuan Zang [1,2], Qun Liu [1,2], Rui Han [1,2,*] and Xu Gong [3]

[1] Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[2] Chongqing Key Laboratory of Computational Intelligence, Chongqing University of Posts and Telecommunications, Chongqing 400065, China
[3] Chongqing Key Laboratory of Toxicology and Drug Analysis, Chongqing Police College, Chongqing 401331, China
* Correspondence: d230201010@stu.cqupt.edu.cn

How To Cite: Zang, M.; Liu, Q.; Han, R.; et al. MG-CMIF: A Multi-Granularity Enhanced Cross-Modal Information Fusion Framework for Molecular Property Prediction. *Journal of Machine Learning and Information Security* **2026**, *2*(3), 18. https://doi.org/10.53941/jmlis.2026.100018

Received: 14 May 2026
Revised: 17 August 2026
Accepted: 31 August 2026
Published: 7 September 2026

Abstract: Molecular property prediction provides an important computational basis for compound screening and drug development by estimating physicochemical characteristics and biological activities from molecular structures. Although deep learning has improved molecular modeling, existing methods often describe molecules through a limited structural view or combine multiple views without sufficiently exploiting their complementary relationships. In addition, graph-based approaches commonly concentrate on local atomic connectivity, making it difficult to represent chemically meaningful structural units that may strongly influence molecular properties. This paper develops MG-CMIF, a multi-granularity cross-modal framework for molecular property prediction. The proposed model describes each molecule from symbolic, topological, and spatial perspectives and learns an integrated representation through three key designs. First, hierarchical graph modeling combines detailed atomic interactions with substructure-level chemical patterns to enrich topology-oriented features. Second, interaction across molecular views enables information relevant to property prediction to be exchanged selectively rather than merged through shallow operations. Third, alignment-oriented training objectives encourage representations derived from the same molecule to preserve compatible chemical semantics during fusion. Experiments on multiple public benchmark datasets show that MG-CMIF achieves better prediction results than competitive methods in both classification and regression settings. Further ablation analyses confirm that hierarchical structural modeling and cross-view integration both contribute to the effectiveness of the proposed framework.

Keywords: molecular property prediction; graph neural network; multi-modal learning; multi-granularity representation; self-supervised learning

1. Introduction

The high cost of experimental drug development makes efficient early-stage molecular screening increasingly important, since bringing a new drug to market generally requires more than one billion US dollars [1]. Throughout the discovery process, molecular properties provide essential information for evaluating, selecting, and generating candidate compounds [2]. The practical value of computational molecular analysis depends heavily on whether a model can encode structure-related chemical characteristics in an informative form, as such knowledge underlies various drug-oriented prediction and design applications [3–6]. Benefiting from advances in deep learning

techniques [7–10], neural models are increasingly used to derive property-relevant features directly from molecular descriptions, thereby providing a more effective basis for molecular property prediction.

Molecular structures can be described from different views, which mainly include sequence-based, graph-based, and geometry-based representations. Sequence-based representations, represented by SMILES [11], transform atoms, chemical bonds, functional groups, and other chemical information into textual strings, but they do not explicitly preserve molecular topology or spatial structure. Graph-based representations model molecules as graphs, where atoms and bonds are encoded as nodes and edges, making them suitable for capturing connectivity patterns and structural fragments such as rings. However, molecular graphs are still limited in describing three-dimensional conformations and stereochemical information such as atomic chirality. Geometry-based representations further introduce spatial information, including atomic coordinates, bond angles, and torsion angles, thereby providing complementary cues for molecular structure modeling. Despite these advantages, each individual representation only reflects molecules from a specific perspective and is insufficient to characterize complex molecular properties comprehensively. Therefore, multi-modal molecular representation learning has received increasing attention. However, many existing methods only combine two types of molecular information, which restricts the completeness of learned representations. Even though several studies incorporate sequence, graph, and geometry simultaneously [12,13], most existing approaches mainly focus on combining heterogeneous molecular views through weighted summation or direct concatenation. Such strategies usually treat each modality as an independent feature source and lack effective interaction between different structural granularities and molecular views. Moreover, semantic consistency among sequence, graph, and geometric representations is rarely explicitly optimized during fusion, which may limit the ability of the model to learn chemically meaningful multi-modal representations.

In addition, many graph neural network-based molecular models still perform representation learning mainly at the atom level. Although atom-level message passing is effective in capturing local connectivity and fine-grained structural patterns, it is less capable of explicitly describing higher-level chemical substructures, such as motifs, functional groups, and ring systems. For many molecular properties, the decisive factors are not limited to individual atomic neighborhoods, but also involve chemically meaningful substructures that remain relatively stable across molecules. For instance, functional groups often have a strong influence on toxicity, solubility, and biological activity, while these substructure-level effects may be weakened or implicitly hidden when the model relies only on atom-wise message propagation. Therefore, introducing motif-level modeling together with atom-level graph learning can provide complementary structural information and improve the expressive ability of molecular graph representations.

To address the above issues and build on previous studies [14], we develop MG-CMIF to learn molecular features through complementary structural views and hierarchical chemical information. The framework receives molecular descriptions from symbolic notation, topological connectivity, and spatial conformation, and processes them using dedicated encoders to preserve the characteristics carried by each view. Within the topology-oriented branch, the molecule is represented at two structural scales: individual atoms retain detailed bonding patterns, whereas extracted motifs describe chemically meaningful fragments. Two parallel MPNNs encode these representations separately, and an adaptive weighting component combines them according to their contributions to the learned graph feature. The resulting graph information is then integrated with the sequence-derived and geometry-derived features through cross-attention, allowing relevant cues from different views to be selectively associated rather than simply accumulated. In addition, MG-CMIF introduces a representation contrast objective together with a modality correspondence objective to encourage semantic agreement among features originating from the same molecule. Through these designs, the framework produces molecular representations that contain both detailed structural evidence and higher-level chemical semantics. The resulting technical advances are outlined below:

- A multi-granularity enhanced cross-modal information fusion framework, namely MG-CMIF, is proposed. It unifies sequence, graph, and geometric modalities within an end-to-end model to learn comprehensive molecular representations.
- A novel multi-granularity graph encoder is designed. It constructs atom-level and motif-level graphs simultaneously, and jointly models fine-grained atomic structures and coarse-grained substructures via dual MPNNs and an adaptive granularity fusion module.
- A cross-attention-based fusion mechanism is introduced. In cooperation with contrastive learning and cross-modal matching tasks, it promotes sufficient information interaction and semantic alignment among heterogeneous modalities.
- Extensive comparison and ablation experiments are conducted on multiple public datasets. The results validate the superior performance of MG-CMIF in both molecular classification and regression tasks.

2. Related Work

2.1. Uni-Modal Molecular Representation Learning

Uni-modal molecular representation learning describes molecules from a single structural view, mainly including symbolic sequences, molecular topology, and spatial geometry. Since SMILES converts molecules into linear chemical strings, sequence-oriented studies usually adapt natural language processing models to extract contextual semantics. Early recurrent architectures combine bidirectional sequence modeling with attention to identify informative SMILES patterns, while subsequent pre-training methods learn transferable sequence representations from large molecular corpora [15–17]. However, a symbolic sequence does not explicitly retain the topological and spatial relationships within a molecule. Topology-centered models therefore formulate molecules as graphs and propagate information through atom-bond connections. Methods based on communicative message passing or multiple graph views improve the modeling of local connectivity and bond-related interactions [18–20]. Nevertheless, graph topology alone remains insufficient for describing conformational information such as atomic coordinates, bond lengths, and bond angles. Geometry-aware representation learning further introduces spatial cues into molecular modeling: GEM incorporates geometric relationships into the encoding process, whereas LineEvo progressively evolves features and positions to capture angular and dihedral information [21,22]. These studies show that different unimodal representations emphasize different molecular characteristics, but each view still has inherent limitations in providing a complete structural description.

2.2. Multi-Modal Molecular Representation Learning

Multi-modal molecular representation learning aims to combine complementary information from different molecular descriptions rather than relying on a single view [12,23]. One research direction integrates sequence semantics with graph topology, where SMILES information and structural connectivity are jointly used to enrich molecular representations or support reconstruction objectives [24,25]. Another direction focuses on the correspondence between two-dimensional graphs and three-dimensional conformations. By introducing self-supervised objectives or masking-based pre-training, these methods learn associations between molecular topology and spatial structure [26,27]. More recent studies extend this idea to three-modal settings, incorporating sequence, graph, and geometric information within the same prediction framework [13,28]. In addition, several studies focus on adaptive fusion and interaction-oriented fusion strategies. For example, MoLA employs layer-wise adaptive fusion to dynamically integrate heterogeneous molecular representations [29], while MCMPP introduces cross-attention to capture complementary information among SMILES, molecular fingerprints, graph structures, and three-dimensional conformations [30]. Although these efforts broaden the information available for molecular modeling, the interaction among modalities is often implemented through direct concatenation or similarly shallow fusion operations. Consequently, modality-specific cues may not be sufficiently exchanged, and semantic consistency across different molecular views is not explicitly guaranteed. This limitation motivates the design of a fusion framework that can support both effective cross-modal interaction and feature alignment.

2.3. Multi-Granularity Molecular Representation Learning

Multi-granularity molecular representation learning aims to characterize molecular information from hierarchical structural levels, enabling the model to capture fine-grained local topology at the atomic level as well as coarse-grained chemical semantics such as motifs, functional groups, and ring structures. In recent years, relevant research has gradually shifted from single atom-based graph modeling to hierarchical and multi-view modeling. MGSSL [31] incorporates motif information as a self-supervised learning objective to strengthen the modeling of molecular substructures. HiMol [32] builds connections among node, motif, and graph levels through a hierarchical self-supervised graph learning framework to refine molecular representations. MultiGran-SMILES [33] integrates SMILES features at atomic, subgraph, and molecular granularities for molecular property prediction. M2Mol [34] further establishes a multi-view and multi-granularity learning paradigm to jointly model multi-scale information from both sequence and graph perspectives. Although existing studies have demonstrated the effectiveness of multi-granularity structural modeling, most hierarchical strategies mainly focus on extracting information within a single modality, such as graph or sequence representations. The interaction between multi-granularity structural information and heterogeneous molecular views remains insufficiently explored. To address this limitation, MG-CMIF introduces dual-scale graph encoding at the atomic and motif levels. By deeply fusing multi-granularity graph representations with sequence and geometric features, the proposed framework is capable of learning more comprehensive and discriminative molecular representations.

3. Methods

In this section, we first formulate the studied problem and then describe the design and functionality of each component in the proposed multi-granularity enhanced cross-modal learning framework, MG-CMIF. As illustrated in Figure 1, the overall architecture of MG-CMIF is composed of four main parts: the modality encoder module, the cross-attention-based feature fusion module, the alignment module, and the prediction module.

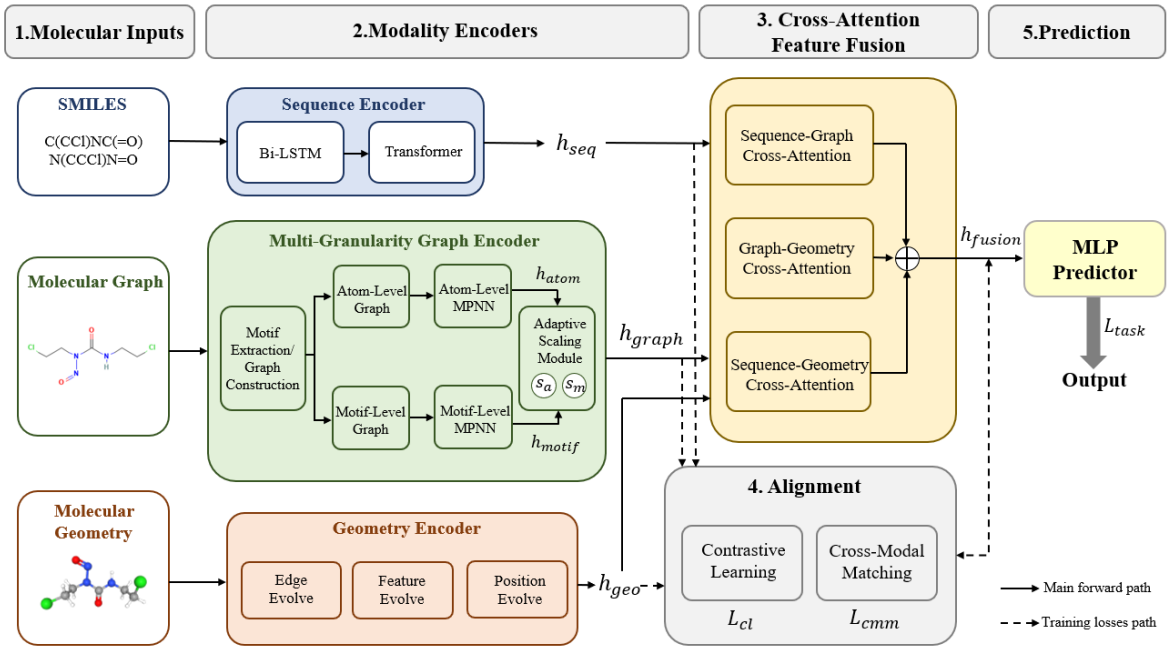

Figure 1. Overview of MG-CMIF.

3.1. Problem Definition

Given a molecule, the purpose of molecular property prediction is to estimate its target physicochemical or biological property, such as solubility or toxicity, from available structural descriptions. This task requires transforming the original molecular information into a representation that preserves property-relevant characteristics. In MG-CMIF, the molecule is described through complementary sequence, graph, and geometric views, which provide symbolic composition, topological connectivity, and spatial conformation information, respectively. Let X denote the encoded multi-modal molecular input. The representation learning process and the subsequent prediction procedure are formulated as follows:

$$H = f(X) \tag{1}$$

$$y = P(H) \tag{2}$$

where f denotes the neural network responsible for mapping the input X into the learned molecular representation H, P represents the property predictor, and y is the predicted molecular property. By incorporating SMILES sequences, molecular graphs, and molecular geometric features into X, the model is expected to provide a more informative basis for property prediction.

3.2. Modality Encoders

The modality encoders of MG-CMIF consist of three independent encoders: a Sequence Encoder, a Multi-Granularity Graph Encoder, and a Geometry Encoder. The detailed introduction of the modality encoders is presented as follows.

3.2.1. Sequence Encoder

SMILES describes a molecule as an ordered sequence of chemical tokens, which provides symbolic information complementary to graph topology and geometric structure. Let a molecular sequence be denoted as $S = s_i | i = 1, \ldots, T$. Each token is first mapped by one-hot encoding to obtain the input feature sequence $X = x_t | t =$

$1, \ldots, T$. A Bi-LSTM is then applied to encode contextual information from both directions, and its forward and backward hidden states are concatenated as the token-level representation h_t:

$$h_t = \text{Concat}\left(\overrightarrow{LSTM}(x_t, \vec{h}_{t-1}), \overleftarrow{LSTM}(x_t, \overleftarrow{h}_{t-1})\right) \quad (3)$$

After all tokens are processed, the sequence representation is formed as $H_{\text{seq}} = \{h_1, \ldots, h_T\}$. Since chemical dependencies in SMILES may occur between tokens separated by relatively long distances, H_{seq} is further passed through the Transformer module to capture long-range relationships within the molecular sequence.

3.2.2. Multi-Granularity Graph Encoder

The multi-granularity graph encoder is designed to learn atom-level molecular graphs with fine-grained structural information while explicitly extracting molecular motifs to build coarse-grained motif-level graphs. Through this design, the model can describe not only atomic interactions but also the connectivity patterns among different motifs during representation learning. For this purpose, Algorithm 1 is introduced to construct motif-level graphs. Specifically, the motif extraction procedure, denoted as the ExtractMotif(·) function, is first used to recognize and extract motif units from each molecule.

Motif Extraction: In the motif extraction process, three categories of substructures are selected as the fundamental units for constructing motif-level graphs, and the detailed rules are defined as follows: (1) Cyclic structures: each ring structure in a molecule is regarded as a separate motif. When two rings contain two or more shared atoms, they are combined into a new motif. (2) Acyclic structures: for the acyclic parts of a molecular graph, functional atoms are first detected and marked, including heteroatoms, carbon atoms involved in non-aromatic double or triple bonds, and acetal carbon atoms. Here, acetal carbon atoms refer to sp^3-hybridized carbon atoms bonded to at least two oxygen, nitrogen, or sulfur atoms. Then, the connected marked atoms and their neighboring ambient carbon atoms are grouped into one motif, where ambient carbon atoms indicate unmarked carbon atoms directly connected to the marked atoms. (3) Carbon-carbon single bonds: for the remaining regions that are not included in the above two types, each uncovered carbon-carbon single bond is taken as an independent motif, thereby ensuring a complete partition of the atom-level molecular graph.

Construction of Motif-level Graph: As shown in Algorithm 1, after all motif nodes are obtained, edges between motif nodes are further constructed according to their structural relationships. (1) When two motifs contain shared atoms, these common atoms are used to define the edge connecting the corresponding motif nodes. (2) When two motifs have no overlapping atoms, all atoms in one motif are traversed to identify atoms that are adjacent to atoms in the other motif. The selected atoms and their adjacent atoms in the paired motif are then used together to establish structural edges between the two motif nodes.

After the above construction process, the motif-level graph can be represented as $G_m = (V_m, E_m)$. Specifically, $V_m = \{n_{m,u} | u \in [1, \ldots, N_m]\}$ refers to the set of motif nodes in a molecule, where each node corresponds to an extracted motif. $E_m = \{e_{m,uv} | u, v \in [1, \ldots, N_m]\}$ represents the edge set that describes the connections between different Motifs, and N_m denotes the number of Motifs contained in the molecule. In the atom-level graph construction, each atom node is initialized with a feature vector containing fundamental chemical attributes of atoms. Specifically, the atom features include atom type, degree, formal charge, hybridization state, hydrogen count, and aromaticity information. The bond features are constructed based on chemical bond properties, including bond type, conjugation information, and ring membership. To describe the internal composition of each Motif, the atomic features and chemical bond features belonging to the corresponding Motif are first accumulated and then concatenated. The obtained vector is used as the initial feature representation of the Motif, which can be written as:

$$x_{m,u} = \sum_{n_{a,i} \in n_{m,u}} x_{a,i} \oplus \sum_{e_{ij} \in n_{m,u}} x_{a,ij} \in R^{d_n + d_e} \quad (4)$$

where $\oplus$ indicates the concatenation operation between feature vectors, $x_{a,i}$ denotes initial atom feature vector, $x_{a,ij}$ represents the initial feature vector of the bond connecting atom nodes, and $N(n_{a,i})$ denotes the set of neighboring nodes of node $n_{a,i}$. For the edges in the motif-level graph, their initial features are generated from the atomic features associated with the corresponding motif-edge connection, which is defined as:

$$x_{m,uv} = \sum_{n_{a,i} \in e_{m,uv}} x_{a,i} \in R^{d_n} \quad (5)$$

Algorithm 1 Construction of Motif-Level Graph

Input: Atom-level graph $G_a = (V_a, E_a)$; ExtractMotif(·)
Output: Motif-level graph $G_m = (V_m, E_m)$
1: Initialize motif node set $V_m \leftarrow \{\}$; 2: Initialize motif edge set $E_m \leftarrow \{\}$;
3: **for** each motif obtained by ExtractMotif(G_a) **do**
4: $V_m \leftarrow V_m \cup \text{Motif}$
5: **end for**
6: **for** $n_{m,u} \in V_m$ **do**
7: **for** $n_{m,v} \in V_m$ **do**
8: **if** $nm,u \cap n_{m,v} \neq \emptyset$ **then**
9: $e_{m,uv} \leftarrow nm,u \cap n_{m,v}$;
10: **end if**
11: **if** $nm,u \cap n_{m,v} = \emptyset$ **then**
12: **for** $na,i \in nm,u$ **do**
13: **if** $na,j \in N(na,i)$ and $na,j \in n_{m,v}$ **then**
14: $e_{m,uv} \leftarrow e_{m,uv} \cup \{na,i,\ na,j\}$;
15: **end if**
16: **end for**
17: **end if**
18: $E_m \leftarrow E_m \cup e_{m,uv}$;
19: **end for**
20: **end for**
21: $G_m = (V_m, E_m)$;
22: **return** G_m;

After obtaining the motif-level graph, a message passing neural network (MPNN) is further employed to learn substructure-level molecular representations, thereby providing complementary information to atom-level molecular features.

MG-CMIF uses the MPNN module to capture molecular representations from different granularities. Specifically, information propagation is first conducted on the atom-level graph and the motif-level graph through Atom-MPNN and Motif-MPNN, respectively. During this process, each node aggregates messages from its neighboring nodes to generate the corresponding message representation. For an atom node na,i and a motif node nm,u, the message aggregation operations of Atom-MPNN and Motif-MPNN at the l-th layer are defined as follows:

$$m_{a,i}^{(l)} = \sum_{n_{a,j} \in N(n_{a,i})} \left(f_a\left(x_{a,ij}\right) \odot h_{a,j}^{(l)} \right) \tag{6}$$

$$m_{m,u}^{(l)} = \sum_{n_{m,v} \in N(n_{m,u})} \left(f_m\left(x_{m,uv}\right) \odot h_{m,v}^{(l)} \right) \tag{7}$$

where $f_a(\cdot)$ and $f_m(\cdot)$ are independent linear layers for atom-level and motif-level representations, respectively. Both mapping functions are implemented as trainable projection layers to transform the original graph representations into a unified dimensional space d, facilitating subsequent granularity fusion. $\odot$ indicates the element-wise multiplication operation. The initial hidden states are given by $h_{a,i}^{(0)} = W_1 x_{a,i} \in R^d$ and $h_{m,u}^{(0)} = W_2 x_{m,u} \in R^d$.

Then, in the node state updating stage, the instantiated MPNN adopts the Gated Recurrent Unit (GRU) [35] to update node hidden states for subsequent message propagation. By controlling the information flow between the aggregated messages and the previous hidden states, GRU helps the model better capture the dependencies involved in representation learning. At the l-th layer, the update processes of Atom-MPNN and Motif-MPNN are expressed as:

$$h_{a,i}^{(l+1)} = \text{GRU}\left(m_{a,i}^{(l)}, h_{a,i}^{(l)}\right) \tag{8}$$

$$h_{m,u}^{(l+1)} = \text{GRU}\left(m_{m,u}^{(l)}, h_{m,u}^{(l)}\right) \tag{9}$$

After L rounds of information propagation, Atom-MPNN and Motif-MPNN respectively apply global pooling (GP) and max pooling (MP) [36] in the readout stage to aggregate node-level representations and obtain graph-level molecular representations. This readout strategy improves the stability and robustness of the learned graph representations.

$$H_a = \mathrm{GP}\left(\{h_{a,i}^{(L)} \mid n_{a,i} \in V_a\}\right) \oplus \mathrm{MP}\left(\{h_{a,i}^{(L)} \mid n_{a,i} \in V_a\}\right) \tag{10}$$

$$H_m = \mathrm{GP}\left(\{h_{m,u}^{(L)} \mid n_{m,u} \in V_m\}\right) \oplus \mathrm{MP}\left(\{h_{m,u}^{(L)} \mid n_{m,u} \in V_m\}\right) \tag{11}$$

The Multi-Granularity Graph Encoder further adopts an adaptive scaling module to adjust graph representation vectors automatically, so that atom-level and motif-level features can be transformed into a more suitable representation space. Specifically, learnable parameter matrices $W_a, W_m \in R^{d\times 1}$ are first used to project and compress the atom representation and motif representation, respectively, thereby generating the corresponding feature description vectors:

$$z_a = W_a H_a, z_m = W_m H_m. \tag{12}$$

where z_a and z_m represent the description vectors obtained from atom-level and motif-level representations, respectively, and d denotes the dimension of the hidden representation space.

Then, the two description vectors are integrated into a compact low-dimensional representation. During this process, the gating mechanism is introduced to perceive contextual information and model the dependency relationship between different granularity representations:

$$S = \sigma\left(W_4 \delta\left(W_3(z_a \oplus z_m)\right)\right) \tag{13}$$

where $S = [s_a, s_m]$ is the scaling coefficient vector used to regulate atom and motif representations, δ denotes the ReLU activation function, $W_3 \in R^{2\times r}$ and $W_4 \in R^{r\times 2}$ are learnable parameter matrices, and r is a hyperparameter. The function σ denotes an activation function, such as sigmoid or softmax, which limits the values of the scaling vector S within the range of (0, 1).

Finally, the original graph representations H_a and H_m are reweighted according to their corresponding scaling coefficients. The adjusted representations H_a' and H_m' are further concatenated to form the final graph representation H_g.

$$H_a' = s_a \cdot H_a, H_m' = s_m \cdot H_m. \tag{14}$$

3.2.3. Geometry Encoder

Topological connections alone cannot fully describe spatial characteristics that are relevant to molecular properties. Therefore, MG-CMIF introduces a geometry encoder to complement the structural information provided by the graph branch. In this module, LineEvo [22] is employed for geometric feature learning because it can be incorporated into general graph neural networks without changing their basic encoding paradigm. Rather than directly operating only on the original atom-bond graph, LineEvo reorganizes molecular geometry through line graph transformation. Each bond in the original graph is mapped to a node in the transformed graph, and two generated nodes are connected when the corresponding bonds share an atom. The initial feature of a line graph node is constructed from the attributes of the two connected atoms and the corresponding bond information, which enables the geometric encoder to preserve both chemical connectivity and local spatial relationships. In this way, local bond relationships are converted into higher-order structural connections. Meanwhile, GATv2 [37] is used to update feature representations, and the spatial position associated with each transformed node is determined by the midpoint of its corresponding atom pair:

$$p_{ij}' = (p_i + p_j)/2 \tag{15}$$

where p_i and p_j denote the coordinates of atom i and atom j, respectively, and p_{ij}' is the evolved position generated from the bond between them.

During geometric representation learning, GATv2 and LineEvo perform complementary operations. Specifically, GATv2 first aggregates neighboring line graph node features through attention-based message passing, allowing each bond representation to incorporate local geometric context. Subsequently, LineEvo updates the line graph structure, node features, and spatial positions based on the evolved geometric relationships. By stacking multiple LineEvo layers, the encoder gradually captures higher-order spatial dependencies, such as bond-angle and dihedral-angle information, which cannot be obtained from the original atom-bond connectivity alone.

3.3. Cross-Attention Feature Fusion

The three molecular modalities encode different but complementary aspects of the same molecule. To enable information from one molecular view to selectively interact with that from another view, MG-CMIF employs a

cross-attention-based fusion mechanism. After the three modality encoders generate the corresponding embeddings, denoted as H_{seq}, H_{graph}, and H_{geo}, pairwise cross-modal interactions are performed among sequence, graph, and geometric representations. In each interaction, the query representation retrieves informative features from the paired modality through key-value matching, allowing the model to emphasize cross-modal dependencies that are more relevant to molecular characterization. The three interaction representations are calculated as follows:

$$\text{Attn}_{\text{graph}-\text{geo}} = \text{CrossAttention}\left(f_Q(H_{\text{geo}}), f_K(H_{\text{graph}}), f_V(H_{\text{graph}})\right) \tag{16}$$

$$\text{Attn}_{\text{graph}-\text{seq}} = \text{CrossAttention}\left(f_Q(H_{\text{graph}}), f_K(H_{\text{seq}}), f_V(H_{\text{seq}})\right) \tag{17}$$

$$\text{Attn}_{\text{geo}-\text{seq}} = \text{CrossAttention}\left(f_Q(H_{\text{seq}}), f_K(H_{\text{geo}}), f_V(H_{\text{geo}})\right) \tag{18}$$

$$\text{CrossAttention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C/\text{d}}}\right) \cdot V \tag{19}$$

where Q K and V denote the query, key, and value vectors, respectively. C is the embedding dimension, d represents the number of attention heads, and f_Q, f_K, and f_V are projection functions for constructing the attention space. Accordingly, $\text{Attn}_{\text{graph-geo}}$, $\text{Attn}_{\text{graph-seq}}$, and $\text{Attn}_{\text{geo-seq}}$ contain the interaction-enhanced information learned from the corresponding modality pairs. Finally, these three representations are concatenated to form the fused molecular representation used by the subsequent prediction module:

$$H_{\text{fusion}} = \text{Concat}\left(\text{Attn}_{\text{graph}-\text{geo}}, \text{Attn}_{\text{graph}-\text{seq}}, \text{Attn}_{\text{geo}-\text{seq}}\right) \tag{20}$$

3.4. Alignment Module

Although cross-attention enables information exchange among different molecular modalities, it does not guarantee that the representations generated from the same molecule are semantically compatible. To improve the semantic consistency of multi-modal features, MG-CMIF introduces two self-supervised objectives, namely contrastive learning and cross-modal matching. The former encourages representations of the same molecule from different modalities to become closer, while the latter identifies whether a group of modal inputs describes the same molecular instance.

3.4.1. Contrastive Loss

For a given molecule, its sequence, graph, and geometric representations should express consistent chemical characteristics despite being obtained from different views. Therefore, contrastive learning is employed to increase the similarity between modal representations belonging to the same molecule and decrease the similarity between representations derived from different molecules. The contrastive objective is defined as follows:

$$L_{CL} = L_{cl}(z_{\text{seq}}, z_{\text{graph}}) + L_{cl}(z_{\text{seq}}, z_{\text{geo}}) + L_{cl}(z_{\text{graph}}, z_{\text{geo}}) \tag{21}$$

$$L_{cl}(a, b) = -\log \frac{\exp(\text{sim}(a_i, b_i)/\text{T})}{\sum_{j=1, i\neq j}^{N} \exp(\text{sim}(a_i, b_j)/\text{T})} \tag{22}$$

where z_{seq}, z_{graph}, and z_{geo} denote the projected representations of H_{seq}, H_{graph}, and H_{geo}, respectively. $\text{sim}(\cdot,\cdot)$ is the similarity function, and T represents the temperature coefficient.

3.4.2. Cross-Modal Matching Loss

Contrastive learning constrains pairwise modal similarity, whereas cross-modal matching (CMM) further examines whether the three modal representations form a consistent molecular sample. During training, unmatched samples are generated by shuffling the sequence, graph, and geometry inputs within each mini-batch. The model is then required to distinguish matched modal combinations from mismatched ones. The CMM objective is formulated as:

$$L_{CMM} = \sum_{(S,G,H)\in B} \text{crossenlropy}\left(y(S, G, H), p(M(h_S, h_G, h_H))\right) \tag{23}$$

where B denotes the shuffled mini-batch, and y indicates whether the three modal inputs are obtained from the same molecule. p is the matching predictor composed of a fully connected layer and a softmax function. M

represents the multi-modal encoder consisting of the Sequence Encoder, the Multi-Granularity Graph Encoder, the Geometry Encoder, and the Cross-Attention Feature Fusion Module.

3.5. Prediction Module

The fused molecular representation H_{fusion} is finally input into the predictor P to generate the molecular property prediction result. According to different task types, corresponding objective functions are adopted for model optimization. Specifically, binary cross-entropy loss is used for classification tasks, while mean squared error loss is employed for regression tasks. The prediction and training objectives are formulated as follows:

$$y_{\text{pred}} = P(H_{\text{fusion}}) \tag{24}$$

$$L_{\text{comb}} = L_{CL} + L_{\text{CMM}} \tag{25}$$

$$L_{\text{cls}} = L_{\text{BCE}}(y_{\text{pred}}, y) + \lambda \cdot L_{\text{comb}} \tag{26}$$

$$L_{\text{reg}} = L_{\text{MSE}}(y_{\text{pred}}, y) + \lambda \cdot L_{\text{comb}} \tag{27}$$

where P is implemented by fully connected layers, and L_{comb} denotes the combined self-supervised loss formed by L_{CL} and L_{CMM}. Since the self-supervised alignment objectives and the supervised prediction objective may have different optimization scales, a balancing coefficient λ is introduced to control their relative contributions, thereby making the overall training process more stable.

4. Experiments

This section evaluates MG-CMIF from predictive performance, component contribution, and representation quality. The benchmark datasets, competing methods, and implementation details are introduced to establish the experimental setting, followed by quantitative comparisons and ablation studies that assess the overall framework and its key designs. Visualization analyses are further conducted to provide an intuitive understanding of the molecular representations learned by the proposed model.

4.1. Benchmark Datasets

Eight publicly available datasets collected from MoleculeNet [38] are selected to examine MG-CMIF under different molecular property prediction settings. Among them, BACE, BBBP, ClinTox, HIV, and SIDER cover single-label or multi-label classification problems, whereas ESOL, FreeSolv, and Lipophilicity involve the prediction of continuous-valued molecular properties. Table 1 reports the main statistics of these datasets, including the number of compounds, the number of prediction tasks, the evaluation metric, and the target molecular property.

Table 1. Statistics of eight molecular property benchmark datasets.

Dataset	Tasks	Compounds	Task Type	Metric	Property
BACE	1	1513	Classification	ROC-AUC	Inhibitory activity against human β-secretase 1
BBBP	1	2039	Classification	ROC-AUC	Blood-brain barrier permeability
ClinTox	2	1478	Classification	ROC-AUC	Clinical toxicity risk
HIV	1	41,127	Classification	ROC-AUC	Anti-HIV activity
SIDER	27	1427	Classification	ROC-AUC	Drug-associated adverse effects
ESOL	1	1128	Regression	RMSE	Aqueous solubility
FreeSolv	1	642	Regression	RMSE	Hydration free energy in aqueous solution
Lipophilicity	1	4200	Regression	RMSE	Octanol/water distribution coefficient

4.2. Baselines

To evaluate the predictive performance of MG-CMIF, nine representative baseline methods are selected for comparison. These methods cover sequence-based, graph-based, geometry-aware, and multi-modal molecular representation learning approaches:

- Seq2SeqFP [39]: An unsupervised approach that employs a gated recurrent unit (GRU) to derive molecular representations from sequential molecular descriptions.
- SMILES Transformer [16]: A sequence-oriented model that pre-trains a Transformer on SMILES strings to generate molecular fingerprints for downstream tasks.
- CMPNN [19]: A communicative message passing network that improves molecular graph encoding by

enhancing information exchange between atom and bond representations.

- GATv2 [37]: A graph attention architecture that introduces dynamic attention weighting to better characterize the relationships among molecular graph nodes.
- GEM [21]: A geometry-enhanced representation learning method that incorporates molecular geometric information to describe spatial structural characteristics.
- GraSeq [25]: A molecular representation learning method that combines graph structures and sequence information for joint molecular modeling.
- GraphMVP [26]: A molecular pre-training approach that learns representations by exploiting both 2D topological structures and 3D geometric views.
- 3D Infomax [40]: A molecular property prediction method that learns informative representations by maximizing mutual information between graph and three-dimensional molecular views.
- FTMMR [23]: A multi-representation prediction framework that integrates sequence, graph, and geometric information through Transformer-based feature interaction.

4.3. Experimental Settings

All datasets are randomly partitioned into training, validation, and test subsets according to an 8:1:1 ratio, so that model training, hyperparameter selection, and final performance evaluation are conducted on separate samples. Classification performance is evaluated using ROC-AUC, which measures the model's ability to distinguish molecules with different labels, while regression performance is assessed by RMSE, reflecting the deviation between predicted and true molecular property values. Each experiment is repeated with five different random seeds, and the average result is reported to obtain a more reliable performance estimate.

For model optimization, Adam is employed as the optimizer, and the training process lasts for 100 epochs with a batch size of 32. The initial learning rate is selected from 1×10^{-3} and 1×10^{-4}, while the maximum learning rate is chosen from 2×10^{-3} and 2×10^{-4}. Regarding the architecture configuration, the hidden dimension of each modality encoder is set to 256 to maintain a consistent feature dimension across sequence, graph, and geometric representations. The Transformer-based sequence encoder adopts four attention heads for modeling dependencies within SMILES sequences. In addition, the loss coefficient λ is set to 0.08 to balance the prediction objective and the auxiliary alignment objectives, and the contrastive temperature coefficient T is fixed at 0.1. All experiments were conducted on an Ubuntu Server 20.04 LTS platform equipped with NVIDIA GeForce RTX 3090 GPUs (NVIDIA Corporation, Santa Clara, CA, USA).

4.4. Comparative Experiments

This study compares MG-CMIF with nine baseline methods on five classification datasets and three regression datasets. The quantitative results are reported in Table 2. In the table, the best results are highlighted in bold, the second-best results are underlined, and "-" denotes that the corresponding result is unavailable for a given method. ROC-AUC is used to evaluate classification performance, where larger values indicate better results, while RMSE is adopted for regression tasks, where smaller values represent higher prediction accuracy. Based on Table 2, the following observations can be made:

(1) From the perspective of overall performance, MG-CMIF obtains the best results on all eight benchmark datasets. For classification tasks, compared with the second-best methods, MG-CMIF improves the ROC-AUC scores on BBBP, BACE, ClinTox, HIV, and SIDER by 3.9%, 5.0%, 6.8%, 3.3%, and 0.3%, respectively, with an average improvement of about 3.9% over the five classification datasets. For regression tasks, the RMSE values on ESOL, FreeSolv, and Lipophilicity are decreased by 12.0%, 18.9%, and 4.6%, respectively, leading to an average error reduction of approximately 11.8%. These results indicate that MG-CMIF has stronger discriminative ability in classification scenarios and can provide more accurate estimation for continuous molecular properties in regression tasks.

(2) Compared with representative multi-modal and cross-modal methods, including GraSeq, GraphMVP, 3D Infomax, and FTMMR, MG-CMIF shows more obvious performance advantages. GraSeq mainly combines SMILES sequences with graph structural information, but does not involve geometric modality modeling or multi-granularity graph perception. GraphMVP and 3D Infomax emphasize the interaction between 2D molecular topology and 3D geometric structure, while the semantic information contained in SMILES sequences is not sufficiently utilized. Although FTMMR also learns molecular representations from multiple modalities, its graph branch still focuses on atom-level representation learning. Different from these methods, MG-CMIF jointly models sequence, graph, and geometric modalities, and introduces a multi-granularity graph encoder to capture both atom-level and motif-level structural information, thus producing more

complete and robust molecular representations. Taking FTMMR as the reference method, MG-CMIF achieves performance improvements of 29.3%, 6.8%, 6.8%, 6.3%, and 0.3% on BBBP, BACE, ClinTox, HIV, and SIDER, respectively, and reduces RMSE by 28.6%, 48.0%, and 5.4% on ESOL, FreeSolv, and Lipophilicity. This comparison further shows that the proposed multi-granularity enhanced cross-modal fusion strategy is more effective than conventional multi-modal integration methods.

(3) MG-CMIF also brings clear improvements on relatively small datasets, such as ClinTox, ESOL, and FreeSolv. Due to the limited number of training samples, models often suffer from insufficient feature learning and unstable generalization on these datasets. However, MG-CMIF still maintains competitive and stable performance under data-limited conditions. This suggests that the proposed framework can effectively extract key molecular characteristics and structural patterns even when the available dataset size is relatively small. One possible reason is that the joint use of sequence, graph, and geometric information provides complementary supervision signals, which helps alleviate the risk of overfitting to limited samples. In addition, the alignment objectives further regularize the representation space, enabling the model to learn more robust molecular embeddings from small-scale datasets.

Table 2. Quantitative comparison of different methods across eight molecular property benchmarks, where ↑ denotes higher values for better performance, ↓ denotes lower values for better performance, bold indicates the best performance, and underline indicates the second best performance.

Method	Classification (ROC-AUC ↑)					Regression (RMSE ↓)		
	BBBP	**BACE**	**ClinTox**	**HIV**	**SIDER**	**ESOL**	**FreeSolv**	**Lipophilicity**
Seq2SeqFP	0.889	0.753	0.904	-	0.553	1.248	2.865	-
SMILES Transformer	0.906	0.739	0.917	0.683	0.561	1.081	2.376	1.169
CMPNN	0.935	<u>0.884</u>	0.894	0.782	0.637	0.845	1.349	0.614
GATv2	0.894	0.849	0.893	-	0.605	<u>0.633</u>	<u>0.948</u>	<u>0.605</u>
GEM	0.725	0.849	0.911	<u>0.806</u>	0.631	0.789	1.815	0.664
GraSeq	<u>0.942</u>	0.763	0.628	0.772	0.579	1.255	2.739	0.648
GraphMVP	0.735	0.857	0.803	0.770	0.639	1.047	1.867	0.769
3D Infomax	0.695	0.785	0.611	0.754	0.528	0.877	2.256	1.226
FTMMR	0.757	0.869	<u>0.923</u>	0.784	<u>0.640</u>	0.780	1.480	0.610
MG-CMIF (ours)	**0.979**	**0.928**	**0.986**	**0.833**	**0.642**	**0.557**	**0.769**	**0.577**

4.5. Ablation Studies

To investigate whether each molecular view contributes useful information to MG-CMIF, we conduct modality-level ablation experiments by removing the sequence, graph, or geometry branch individually. The corresponding variants are denoted as w/o Sequence, w/o Graph, and w/o Geometry, and their results are reported in Table 3. Across all eight benchmark datasets, the complete model consistently achieves better performance than the three ablated variants. This finding indicates that the three input views are jointly required for effective molecular property prediction: SMILES sequences describe chemical information in symbolic form, molecular graphs encode structural connectivity, and geometric representations provide spatial configuration cues. Their combination enables MG-CMIF to form a more informative molecular representation than any configuration lacking one of these views. The results also reveal that different prediction tasks exhibit different sensitivities to molecular modalities.

In particular, removing the graph branch causes pronounced performance degradation on BACE, FreeSolv, and Lipophilicity, showing that these tasks depend strongly on topology- and substructure-related information captured by the multi-granularity graph encoder. By comparison, the decline on ClinTox is more evident when the sequence branch is excluded, implying that sequence-level chemical patterns are important for toxicity-related prediction. Moreover, the w/o Geometry variant performs worse than the complete model on all datasets, although the decrease is generally smaller, suggesting that spatial conformational features act as stable supplementary evidence rather than serving as the dominant information source in most tasks. Relative evaluation shows that removing geometric features yields an average performance decline of 2.0% in classification ROC-AUC and 3.5% in regression RMSE, indicating that geometric information contributes relatively more substantially to regression tasks. Therefore, Table 3 not only confirms the necessity of multi-modal modeling, but also indicates that molecular properties differ in their dependence on sequence, topological, and geometric information.

Table 3. Ablation analysis of modality contributions across eight molecular property benchmarks, where ↑ denotes higher values for better performance, ↓ denotes lower values for better performance, and bold indicates the best performance.

Method	Classification (ROC-AUC ↑)					Regression (RMSE ↓)		
	BBBP	BACE	ClinTox	HIV	SIDER	ESOL	FreeSolv	Lipophilicity
-w/o Sequence	0.969	0.887	0.905	0.813	0.612	0.573	0.915	0.598
-w/o Graph	0.944	0.712	0.952	0.806	0.581	0.658	1.787	1.165
-w/o Geometry	0.969	0.905	0.973	0.809	0.628	0.584	0.791	0.593
MG-CMIF (ours)	**0.979**	**0.928**	**0.986**	**0.833**	**0.642**	**0.557**	**0.769**	**0.577**

To further evaluate whether the multi-granularity graph encoder can effectively model molecular structures at both fine-grained and coarse-grained levels, we perform a targeted ablation study on this component. Specifically, the multi-granularity graph encoder in MG-CMIF is replaced with a standard MPNN, allowing the model to learn graph representations only from a single granularity. The comparison results are presented in Figure 2. The complete MG-CMIF consistently performs better than MG-CMIF (MPNN) on both types of prediction tasks. Specifically, the ROC-AUC scores are improved by approximately 2.4%, 3.0%, 2.4%, 0.7%, and 3.7% on BBBP, BACE, ClinTox, HIV, and SIDER, respectively. For ESOL, FreeSolv, and Lipophilicity, the corresponding RMSE values decrease by about 2.8%, 2.8%, and 4.8%. The gains observed across classification and regression benchmarks indicate that multi-granularity graph modeling provides broadly useful structural information rather than benefits limited to a particular task type or evaluation metric.

The main reason is that a standard MPNN mainly performs message passing on atom-level molecular graphs, which enables it to capture local topological patterns but makes it difficult to explicitly describe higher-level chemical substructures. In molecular property prediction, many properties are not only determined by local atom-bond connections, but are also closely related to functional groups, ring systems, and molecular motifs with stable chemical semantics. Therefore, relying only on single-granularity atom-level representations may lead to incomplete structural characterization, especially when property-related cues are distributed across larger substructures rather than individual atoms or bonds. This limitation can weaken the model's ability to distinguish molecules with similar local topology but different functional substructures. By contrast, the multi-granularity graph encoder in MG-CMIF jointly learns fine-grained atom-level details and coarse-grained motif-level structural information, allowing the graph branch to provide richer and more discriminative representations for the subsequent cross-modal fusion module. This explains why MG-CMIF achieves more consistent prediction performance than MG-CMIF (MPNN) on both classification and regression datasets.

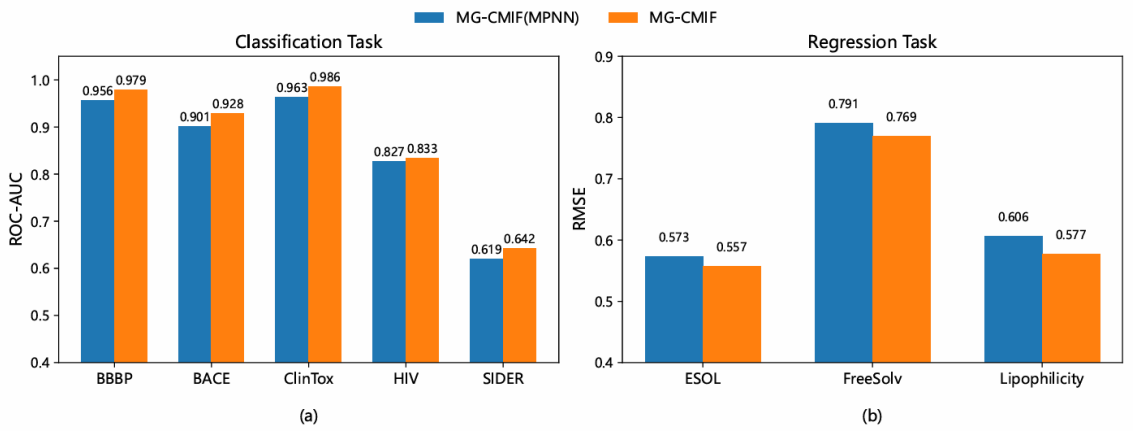

Figure 2. Ablation results of the Multi-Granularity Graph Encoder. (**a**) Classification tasks; (**b**) Regression tasks.

In addition, we investigate the effects of removing the cross-attention feature fusion module and the alignment module, which are denoted as w/o cross-attention and w/o alignment, respectively. In the former variant, cross-attention is replaced by simple feature concatenation, while the latter removes the auxiliary alignment objectives. The corresponding results are reported in Figure 3. Compared with both variants, the complete MG-CMIF obtains better results on the five classification datasets, indicating that feature interaction and representation consistency are both important for multi-modal molecular property prediction.

Specifically, simple concatenation only combines different modality features at the vector level, but it cannot sufficiently model the dependency relationships and complementary information among sequence, graph, and geometric representations. Such a shallow fusion strategy may also introduce redundant or weakly related features,

which can limit the discriminative ability of the final molecular representation. In contrast, the cross-attention mechanism enables different modalities to interact with each other during feature fusion, allowing the model to capture more informative cross-modal associations. Meanwhile, the alignment module introduces contrastive learning and cross-modal matching to reduce semantic gaps among heterogeneous modalities, which helps the fused representation become more consistent and discriminative. By encouraging representations of the same molecule from different modalities to be closer in the latent space, the alignment module further improves the reliability of multi-modal fusion. Therefore, the performance gains shown in Figure 3 demonstrate that cross-modal interaction and semantic alignment are both necessary for improving the prediction ability of MG-CMIF.

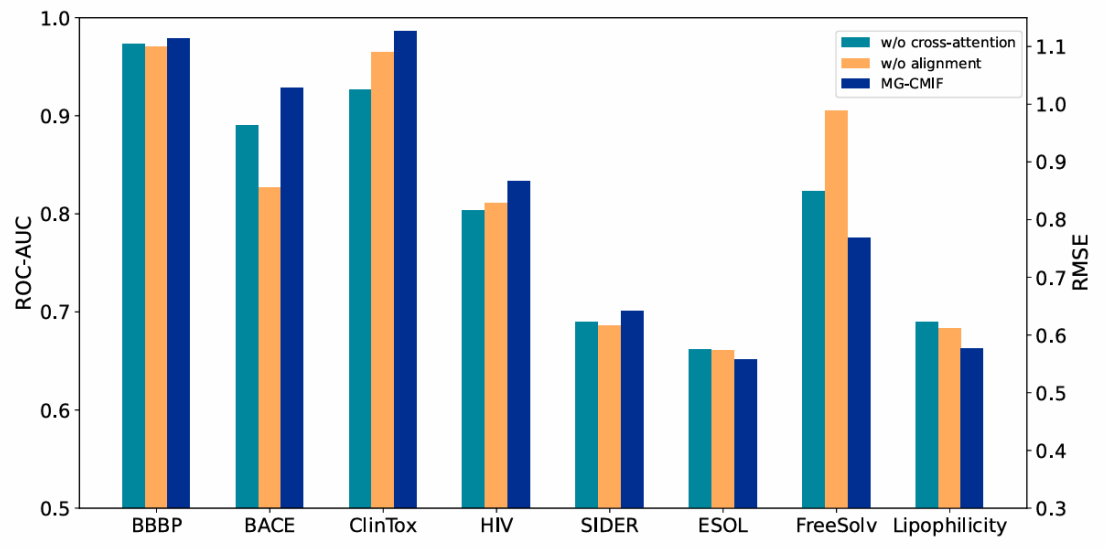

Figure 3. Effect of cross-modal interaction and alignment supervision on molecular property prediction performance.

4.6. Visualization Experiments

4.6.1. Molecular Representation Visualization

To provide an intuitive illustration of MG-CMIF in integrating multi-modal molecular information, t-SNE [41] is used to reduce the dimensionality of molecular representations learned on the BBBP dataset and visualize their distribution, where the t-SNE dimensionality reduction algorithm is implemented using scikit-learn (version 1.3.2). The representations generated by the three single-modality encoders are also compared with those learned by MG-CMIF. As shown in Figure 4, the multi-granularity graph encoder and MG-CMIF can both separate positive and negative samples relatively clearly. More notably, MG-CMIF exhibits clearer class separation and more compact clusters, indicating that it can learn more complete molecular representations. The sequence encoder, which is based on SMILES strings and mainly focuses on global semantic information, does not produce sufficiently discriminative representations for this task. The geometry encoder shows less distinct separation between positive and negative samples in the two-dimensional t-SNE projection, suggesting that geometric information alone is not enough to describe property-related molecular features and should be combined with 2D topological structures and 1D sequence information. Even so, the geometry encoder still shows a certain ability to cluster molecules with different physicochemical characteristics, which reflects its usefulness in capturing structural variations among molecules. Overall, by jointly incorporating sequence, graph, and geometric information, MG-CMIF obtains more discriminative molecular representations and achieves better performance in molecular property prediction.

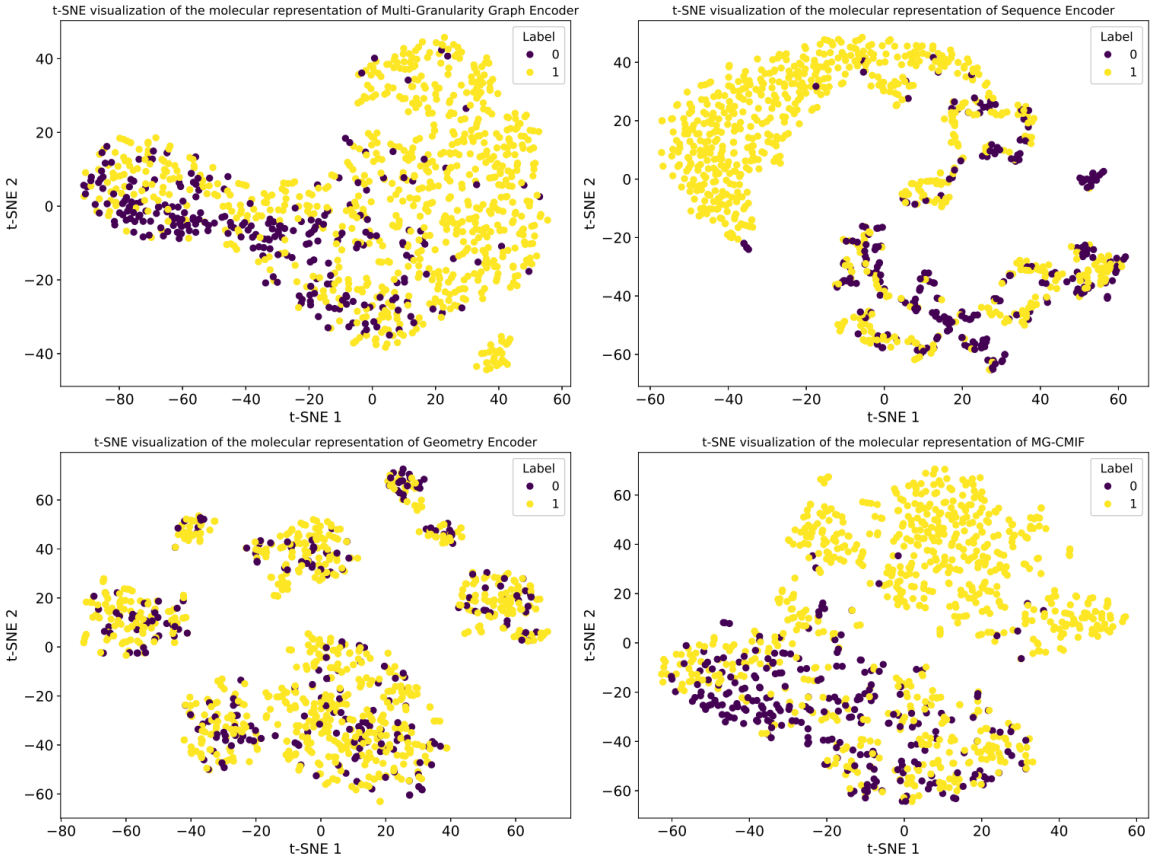

Figure 4. t-SNE projection of molecular representations from different modalities on the BBBP dataset.

4.6.2. Attention Score Visualization

To further investigate the role of the alignment module in MG-CMIF, several representative molecules are selected from the BACE dataset, and their attention scores are visualized using XSMILES (version 0.7.0) [42]. As shown in Figure 5, the model exhibits different attention patterns before and after introducing the alignment module. Without the alignment module, the attention scores are distributed over multiple molecular substructures, suggesting that the model has difficulty identifying the most informative regions from molecular representations. By contrast, when the alignment module is incorporated, the attention weights are more concentrated on specific chemical substructures, especially nitrogen-containing functional groups. Nitrogen is a key element in many organic compounds and often appears in the core ring scaffolds of drug-like molecules. Therefore, assigning closer attention to nitrogen-related substructures can provide useful clues for molecular property prediction. These results indicate that the combined loss function helps guide the model to focus on important molecular substructures during cross-modal feature fusion.

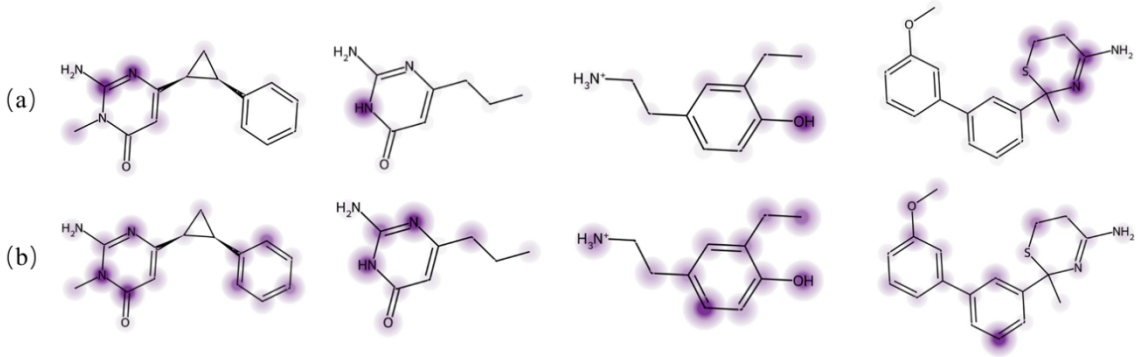

Figure 5. Visualization and comparison of attention scores on four molecules with and without the alignment module. (**a**) With alignment module; (**b**) Without alignment module.

In addition, attention heat maps are used to visualize the attention weight assigned to each atom in the selected molecules, providing a more direct view of the atomic groups that contribute to molecular properties. As shown in

Figure 6, several atomic groups receive relatively high attention weights, indicating that the model assigns greater importance to these regions during molecular representation learning.

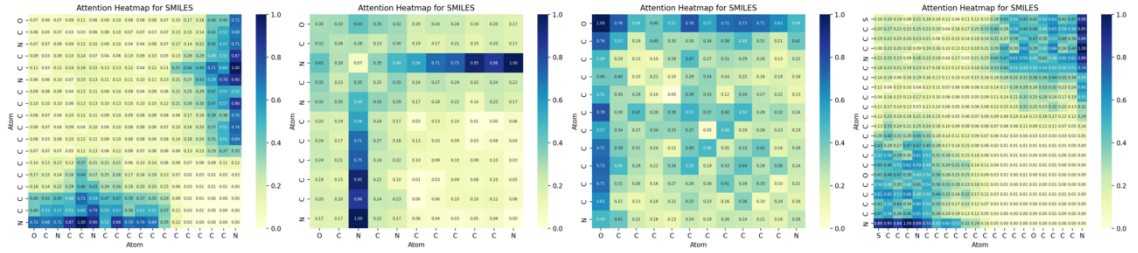

Figure 6. Attention weight heatmaps on four molecules.

4.6.3. Molecular Similarity Visualization

To examine whether the molecular embeddings learned by MG-CMIF are consistent with chemical domain knowledge, four pairs of molecules with similar scaffolds within each pair but distinct scaffolds across different pairs are randomly selected from the BACE dataset. These molecular structures are shown in Figure 7, where each pair contains molecules with either similar scaffolds or clearly different skeleton structures. As presented in Figure 8, the trained model is used to generate embedding vectors for the eight molecules, and cosine similarity is then calculated between different molecular embeddings to assess the model's ability to distinguish molecules with diverse scaffolds. The results show that molecules within the same pair obtain relatively high cosine similarity, while their similarities with molecules from other pairs remain low. This phenomenon indicates that MG-CMIF can preserve meaningful structural relationships in the learned embedding space, consistent with established chemical knowledge. Molecules with similar scaffolds tend to be mapped to closer regions, suggesting that the model captures shared substructure patterns and scaffold-level chemical semantics. In contrast, molecules with obvious structural differences are assigned lower similarity scores, which reflects the discriminative ability of the learned representations. Therefore, the molecular similarity visualization further demonstrates that MG-CMIF can learn chemically meaningful and structurally aware molecular representations, and the learned embeddings are well aligned with practical chemical domain knowledge.

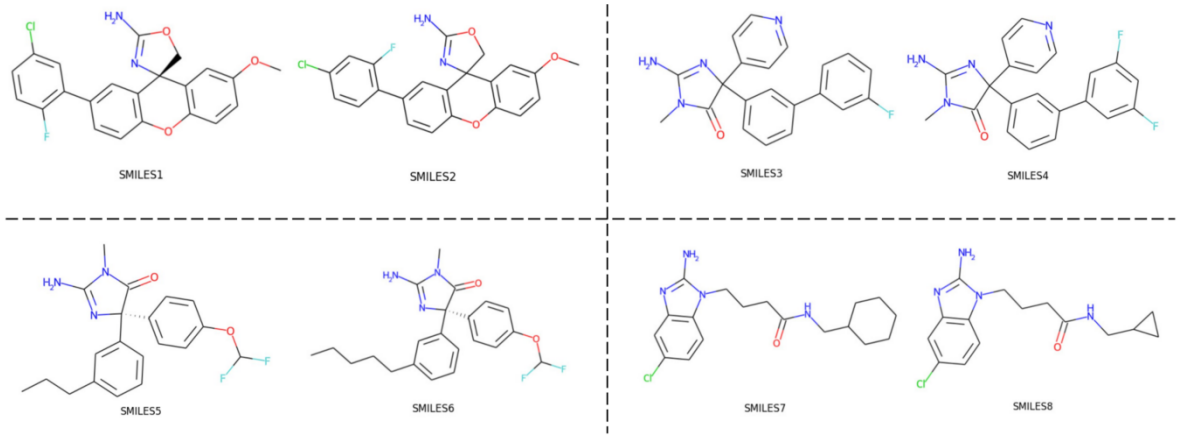

Figure 7. Molecular pairs with similar scaffolds.

Figure 8. Similarity heatmap.

5. Conclusions

This paper presents MG-CMIF, a molecular property prediction framework that integrates sequence semantics, graph structures, and spatial conformations within a unified multi-modal learning architecture, while incorporating hierarchical structural modeling and consistency-oriented fusion to obtain informative molecular representations. Experiments on multiple benchmark datasets show that multi-granularity graph representations improve the characterization of chemically meaningful structural patterns, and that cross-modal interaction together with alignment constraints contributes to more accurate and stable predictions in both classification and regression tasks.

Although MG-CMIF introduces additional computational overhead compared with single-modal molecular representation methods due to the adoption of multiple modality encoders, cross-modal interaction modules, and auxiliary alignment objectives, these additional components effectively capture complementary molecular information from diverse structural perspectives and improve representation quality. Future work will include a comprehensive efficiency evaluation of training cost, memory consumption, and inference speed compared with existing multimodal approaches, as well as the exploration of refined strategies for mining property-relevant molecular substructures and adaptive mechanisms to dynamically select optimal modality and granularity information for diverse prediction tasks.

Author Contributions

M.Z.: Conceptualization, methodology, software, validation, data curation, visualization, writing—original draft preparation. Q.L.: Conceptualization, supervision, resources, project administration, writing—review and editing. R.H.: Conceptualization, methodology, investigation, validation, and writing—review and editing. X.G.: Conceptualization, supervision, methodology, resources, and project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (62576063).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

All data in this study are available from the corresponding author upon reasonable request.

Acknowledgments

The authors would like to thank all individuals and organizations that provided valuable support and assistance during this research. We appreciate the helpful discussions and technical support that contributed to the development and evaluation of the proposed framework.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used AI assistants (ChatGPT 5.6 Sol) to polish manuscript language and assist with the translation of technical terms. After using this service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33.
2. Liu, M.; Yang, Y.; Liu, Q.; et al. A Knowledge-Driven Self-Supervised Approach for Molecular Generation. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2024**, *21*, 1579–1590.
3. Ocaña, A.; Pandiella, A.; Privat, C.; et al. Integrating Artificial Intelligence in Drug Discovery and Early Drug Development: A Transformative Approach. *Biomark. Res.* **2025**, *13*, 45.
4. Liao, Q.; Zhang, Y.; Chu, Y.; et al. Application of Artificial Intelligence in Drug-Target Interactions Prediction: A Review. *npj Biomed. Innov.* **2025**, *2*, 1.
5. Xia, Y.; Xiong, A.; Zhang, Z.; et al. A Comprehensive Review of Deep Learning-Based Approaches for Drug-Drug Interaction Prediction. *Brief. Funct. Genomics* **2025**, *24*, elae052.
6. Gong, X.; Liu, Q.; Han, R.; et al. MIFS: An Adaptive Multipath Information Fused Self-Supervised Framework for Drug Discovery. *Neural Netw.* **2025**, *184*, 107088.
7. Shao, Y.; Li, Y.; Wu, B. A Directional Attention Fusion and Multi-Head Spatial-Channel Attention Network for Facial Expression Recognition. *J. Mach. Learn. Inf. Secur.* **2025**, *1*, 7.
8. Ye, W.; Li, J.; Cai, X. MfGNN: Multi-Scale Feature-Attentive Graph Neural Networks for Molecular Property Prediction. *J. Comput. Chem.* **2025**, *46*, e70011.
9. Peng, J.; Wang, W.; He, Y.; et al. An Enhanced YOLOv8 Algorithm with C2f Parnet for Vehicle and Pedestrian Detection. *J. Mach. Learn. Inf. Secur.* **2026**, *2*, 17.
10. Li, L.; Zhang, Y.; Wang, G.; et al. Kolmogorov-Arnold Graph Neural Networks for Molecular Property Prediction. *Nat. Mach. Intell.* **2025**, *7*, 1346–1354.
11. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
12. Wang, Z.; Jiang, T.; Wang, J.; et al. Multi-Modal Representation Learning for Molecular Property Prediction: Sequence, Graph, Geometry. *arXiv* **2024**, arXiv:2401.03369.
13. Zhang, H.; Wu, J.; Liu, S.; et al. A Pre-Trained Multi-Representation Fusion Network for Molecular Property Prediction. *Inf. Fusion* **2024**, *103*, 102092.
14. Zang, M.; Gong, X.; Han, R.; et al. CMIF: A Cross-Modal Information Fusion Approach for Molecular Property Prediction. In Proceedings of the International Joint Conference on Rough Sets, Chongqing, China, 11–13 May 2025; pp. 228–241.
15. Zheng, S.; Yan, X.; Yang, Y.; et al. Identifying Structure-Property Relationships through SMILES Syntax Analysis with Self-Attention Mechanism. *J. Chem. Inf. Model.* **2019**, *59*, 914–923.
16. Honda, S.; Shi, S.; Ueda, H.R. SMILES Transformer: Pre-Trained Molecular Fingerprint for Low Data Drug Discovery. *arXiv* **2019**, arXiv:1911.04738.
17. Wang, S.; Guo, Y.; Wang, Y.; et al. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Niagara Falls, NY, USA, 7–10 September 2019; pp. 429–436.
18. Li, Z.; Jiang, M.; Wang, S.; et al. Deep Learning Methods for Molecular Representation and Property Prediction. *Drug Discov. Today* **2022**, *27*, 103373.
19. Song, Y.; Zheng, S.; Niu, Z.; et al. Communicative Representation Learning on Attributed Molecular Graphs. In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, Yokohama, Japan, 7–15 January 2021; pp. 2831–2838.

20. Ma, H.; Bian, Y.; Rong, Y.; et al. Cross-Dependent Graph Neural Networks for Molecular Property Prediction. *Bioinformatics* **2022**, *38*, 2003–2009.
21. Fang, X.; Liu, L.; Lei, J.; et al. Geometry-Enhanced Molecular Representation Learning for Property Prediction. *Nat. Mach. Intell.* **2022**, *4*, 127–134.
22. Ren, G.P.; Wu, K.J.; He, Y. Enhancing Molecular Representations via Graph Transformation Layers. *J. Chem. Inf. Model.* **2023**, *63*, 2679–2688.
23. Son, Y.H.; Shin, D.H.; Kam, T.E. FTMMR: Fusion Transformer for Integrating Multiple Molecular Representations. *IEEE J. Biomed. Health Inform.* **2024**, *28*, 4361–4372.
24. Wu, T.; Tang, Y.; Sun, Q.; et al. Molecular Joint Representation Learning via Multi-Modal Information of SMILES and Graphs. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2023**, *20*, 3044–3055.
25. Guo, Z.; Yu, W.; Zhang, C.; et al. GraSeq: Graph and Sequence Fusion Learning for Molecular Property Prediction. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Ireland, 19–23 October 2020; pp. 435–443.
26. Liu, S.; Wang, H.; Liu, W.; et al. Pre-Training Molecular Graph Representation with 3D Geometry. In Proceedings of the Tenth International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
27. Zhu, J.; Xia, Y.; Wu, L.; et al. Unified 2D and 3D Pre-Training of Molecular Representations. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; pp. 2626–2636.
28. Tang, X.; Tran, A.; Tan, J.; et al. MolLM: A Unified Language Model for Integrating Biomedical Text with 2D and 3D Molecular Representations. *Bioinformatics* **2024**, *40* (Suppl. 1), i357–i368.
29. Li, J.; Zhang, Z.; Lei, Z.; et al. MoLA: Molecular Multimodal Layerwise Adaptive Network for Molecular Property Prediction. *Knowl. Based Syst.* **2026**, *338*, 115563.
30. Sun, S.; Wang, P.; He, Y.; et al. Multimodal Cross-Attention Molecular Property Prediction for Text, Sequence, Graph, and Geometry. *ACS Omega* **2025**, *10*, 56225–56239.
31. Zhang, Z.; Liu, Q.; Wang, H.; et al. Motif-Based Graph Self-Supervised Learning for Molecular Property Prediction. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS 2021), Virtual Event, 6–14 December 2021; pp. 15870–15882.
32. Zang, X.; Zhao, X.; Tang, B. Hierarchical Molecular Graph Self-Supervised Learning for Property Prediction. *Commun. Chem.* **2023**, *6*, 34.
33. Jiang, J.; Zhang, R.; Zhao, Z.; et al. MultiGran-SMILES: Multi-Granularity SMILES Learning for Molecular Property Prediction. *Bioinformatics* **2022**, *38*, 4573–4580.
34. Zhang, R.; Wang, X.; Liu, K.; et al. M2Mol: Multi-View Multi-Granularity Molecular Representation Learning for Property Prediction. In Proceedings of the 29th International Conference on Database Systems for Advanced Applications, Gifu, Japan, 2–5 July 2024; pp. 264–274.
35. Cho, K.; van Merriënboer, B.; Gulcehre, C.; et al. Learning Phrase Representations Using RNN Encoder–Decoder for Statistical Machine Translation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
36. Lee, J.; Lee, I.; Kang, J. Self-Attention Graph Pooling. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 3734–3743.
37. Brody, S.; Alon, U.; Yahav, E. How Attentive Are Graph Attention Networks? In Proceedings of the Tenth International Conference on Learning Representations, Virtual Event, 25–29 April 2022.
38. Wu, Z.; Ramsundar, B.; Feinberg, E.N.; et al. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513–530.
39. Xu, Z.; Wang, S.; Zhu, F.; et al. Seq2seq Fingerprint: An Unsupervised Deep Molecular Embedding for Drug Discovery. In Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, Boston, MA, USA, 20–23 August 2017; pp. 285–294.
40. Stärk, H.; Beaini, D.; Corso, G.; et al. 3D Infomax Improves GNNs for Molecular Property Prediction. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 20479–20502.
41. van der Maaten, L.; Hinton, G. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
42. Heberle, H.; Zhao, L.; Schmidt, S.; et al. XSMILES: Interactive Visualization for Molecules, SMILES and XAI Attribution Scores. *J. Cheminform.* **2023**, *15*, 2.